Data in Brief 5 (2015) 605–615

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Data in support of UbSRD: The Ubiquitin Structural Relational Database

Joseph S. Harrison [a,b,*], Tim M. Jacobs [a], Kevin Houlihan [a], Koenraad Van Doorslaer [c], Brian Kuhlman [a,b]

[a] Department of Biochemistry & Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States
[b] Lineberger Comprehensive Cancer Center, Univer sity of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[c] DNA Tumor Virus Section, Laboratory of Viral Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 209892, United States

ARTICLE INFO

Article history:
Received 22 September 2015
Received in revised form
7 October 2015
Accepted 7 October 2015
Available online 19 October 2015

ABSTRACT

This article provides information to support the database article titled "UbSRD: The Ubiquitin Structural Relational Database" (Harrison et al., 2015) [1] . The ubiquitin-like homology fold (UBL) represents a large family that encompasses both post-translational modifications, like ubiquitin (UBQ) and SUMO, and functional domains on many biologically important proteins like Parkin, UHRF1 (ubiquitin-like with PDB and RING finger domains-1), and Usp7 (ubiquitin-specific protease-7) (Zhang et al., 2015; Rothbart et al., 2013; Burroughs et al., 2012; Wauer et al., 2015) [2–5]. The UBL domain can participate in several unique protein–protein interactions (PPI) since protein adducts can be attached to and removed from amino groups of lysine side chains and the N-terminus of proteins. Given the biological significance of UBL domains, many have been characterized with high-resolution techniques, and for UBQ and SUMO, many protein complexes have been characterized. We identified all the UBL domains in the PDB and created a relational database called UbSRD (Ubiquitin Structural Relational Database) by using structural analysis tools in the Rosetta (Leaver et al., 2013; O'Meara et al., 2015; Leaver-fay et al., 2011) [1,6–8]. Querying UbSRD permitted us to report many

DOI of original article: http://dx.doi.org/10.1016/j.jmb.2015.09.011
* Corresponding author.
E-mail address: joseph_harrison@med.unc.edu (J.S. Harrison).

http://dx.doi.org/10.1016/j.dib.2015.10.007

quantitative properties of UBQ and SUMO recognition at different types interfaces (noncovalent: NC, conjugated: CJ, and deubiquitanse: DB). In this data article, we report the average number of non-UBL neighbors, secondary structure of interacting motifs, and the type of inter-molecular hydrogen bonds for each residue of UBQ and SUMO. Additionally, we used PROMALS3D to generate a multiple sequence alignment used to construct a phylogram for the entire set of UBLs (Pei and Grishin, 2014) [9]. The data described here will be generally useful to scientists studying the molecular basis for recognition of UBQ or SUMO.

Specifications Table

Subject area	*Bioinformatics and Biology*
More specific subject area	*Ubiquitin-like homology domain structural biology*
Type of data	*Histograms of per residue properties for UBQ and SUMO, phylogenetic clustering, and UBL schematic.*
How data was acquired	*Computational analysis of protein structures using the Rosetta features analysis protocol*
Data format	*Figures and sqlite3 database*
Experimental factors	*Rosetta3 features analysis of renumbered PDBs*
Experimental features	*We identified all the UBL-containing structures in the PDB, grouped them by type of PPI, and used structural classification tools in Rosetta to quantify measurable properties of these structures.*
Data source location	*University of North Carolina*
Data accessibility	*http://rosettadesign.med.unc.edu/ubsrd/*

Value of the data:

- A description of how we created UbSRD that can be used as a template for researching wishing to construct a Rosetta features database.
- Presents phylogenetic clustering for the ubiquitin homology folds.
- Reports per residue statistics of the molecular properties of UBQ and SUMO participating in protein–protein interactions, generally useful for researchers investigating proteins that recognize UBQ and SUMO.

1. Data experimental design, materials and methods

1.1. Experimental design

1.1.1. Identifying ubiquitin-homology domains in the PDB and constructing an Rosetta features SQL database

To identify all the all the UBL domains in the PDB, we used delta PSI-blast since the standard blast algorithm produced many false positives [10,11]. Using the sequences of UBQ, SUMO and SMT3, the *S.*

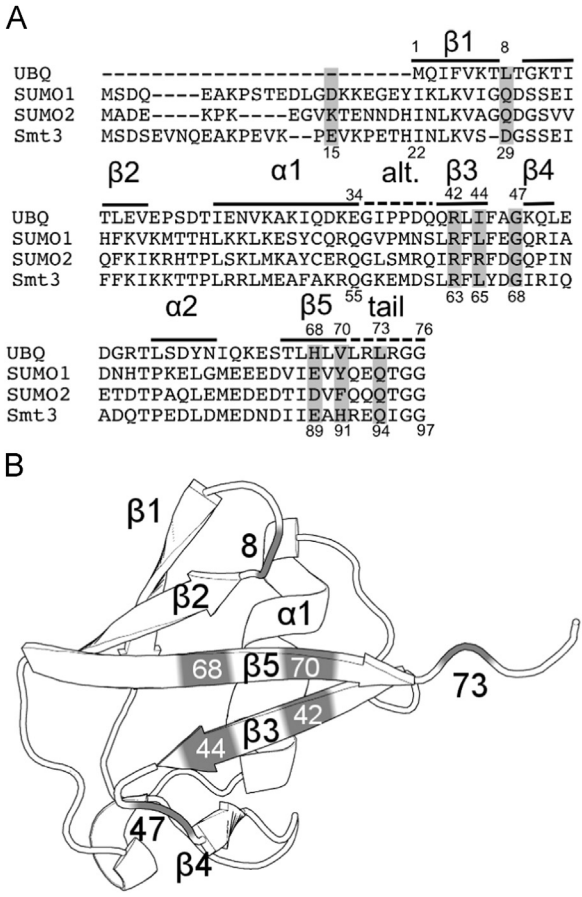

Fig. 1. (A) Sequence alignment between ubiquitin (UBQ), SUMO1, SUMO2 and SMT3, the *S. cerevisiae* SUMO homolog. (B) Cartoon representation of the ubiquitin-like homology fold (UBL) with secondary structure elements annotated.

cerevisiae SUMO homolog, we performed seven iterative rounds of delta-psi blast, downloaded the hit table, and used a one line shell script { grep -o "pdb|\ < ....\ > |" "hit_table_file_name" | cut -d'|' -f2 | sort | uniq } to generate a list of PDB codes to run the Rosetta features analysis on [6]. The features analysis is invoked through the Rosetta Scripts interface and the executable, flags, and Rosetta script needed to run this analysis are found in Supporting file 1 [12]. This analysis will create and SQLite database of Rosetta derived features (for SQLite syntax see [13]) and the recorded features in UbSRD are listed in the Rosetta Script Supporting file 1. It is worth noting the importance of the "jd2:delete_old_poses" flag when running this analysis, otherwise each structure will be stored in memory using a lot of RAM. We manually categorized each structure by the type of UBL and PPI and then used a series of Python scripts to identify, renumber, and generate an SQL table of UBQ and SUMO chains [1]. We further classified each UBQ and SUMO chain by the type of PPI and for UBQ, the type of polymer. Each manually generated table was imported into the SQL database and the syntax for creating and importing tables into existing SQL databases is found in Supporting file 2. We employed a 6 Å distance cutoff from the action coordinate, the average geometric center of the side chain, as a criterion for selecting neighboring residues and the SQL query used to report the residue neighbors is

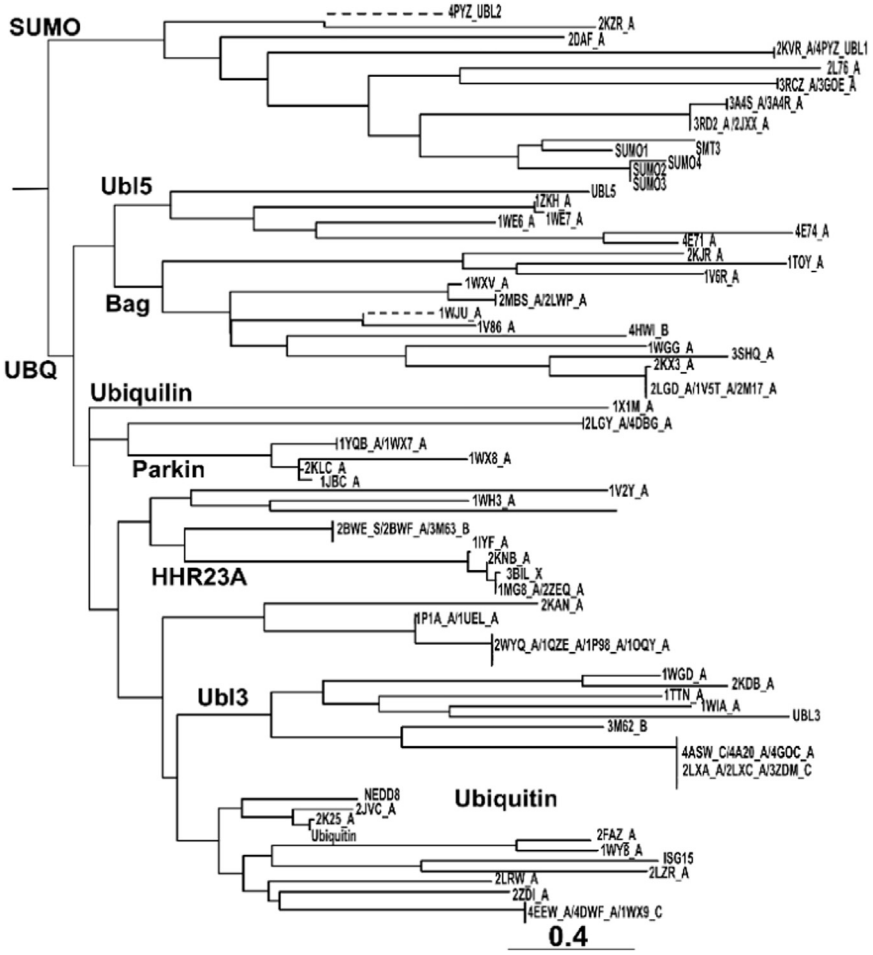

Fig. 2. (A) Phylogenetic clustering of UBLs in UbSRD, dashed lines indicate longer branches (see [1] for methods). An expanded version of the phylogram can be found at http://rosettadesign.med.unc.edu/ubsrd/browse/phylogeny.

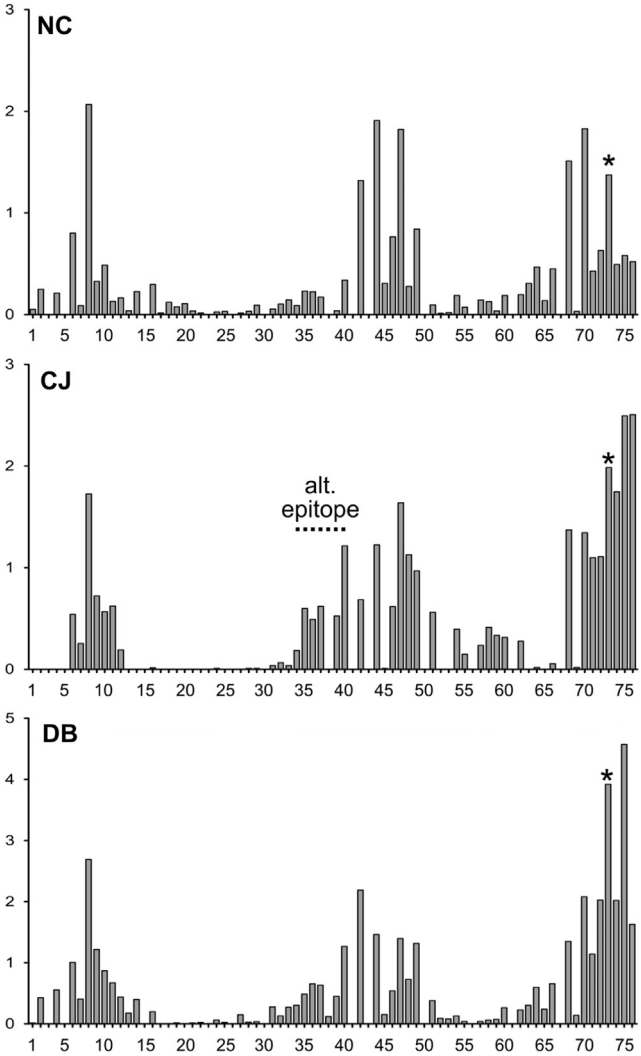

Fig. 3. Number of non-UBQ amino acid neighbors for each residue of UBQ using a 6 Å distance cutoff. The ubiquitin structures are grouped by the following protein–protein interactions: NC: noncovalent, CJ: conjugated, DB: deubiquitinase. The units on the *Y*-axis are average number of normalized non-UBQ neighboring residues per PDB.

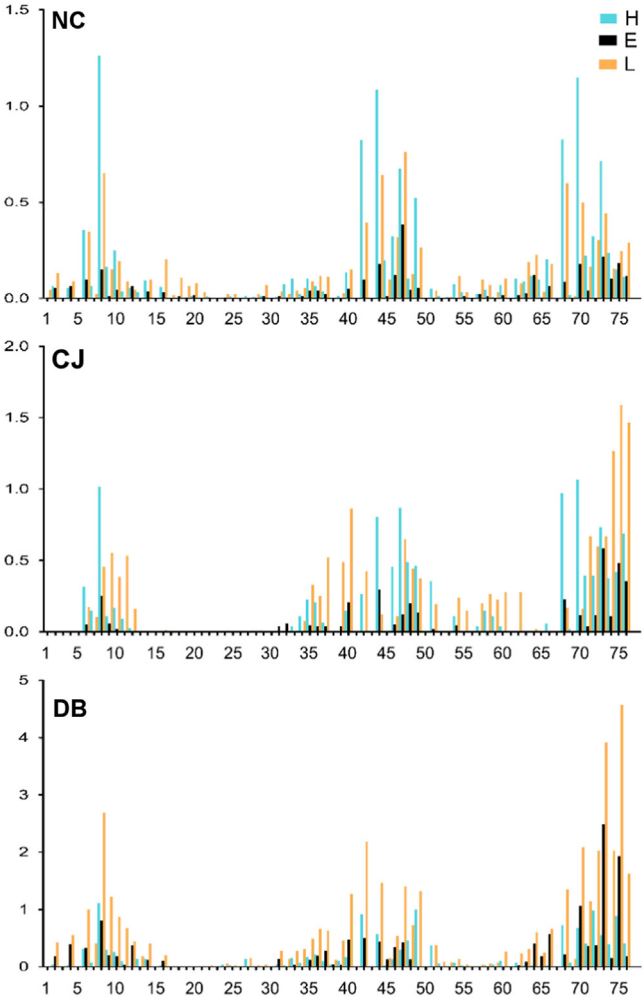

Fig. 4. Secondary structure of UBQ interacting motifs for each residue of UBQ. We classified secondary structure using the simplified DSSP distinction, *H α-helix*, *E β-strand*, *L loop*. The ubiquitin structures are grouped by the following protein–protein interactions: NC: noncovalent, CJ: conjugated, DB deubiquitinase. The *Y*-axis represents the average number of normalized interacting from each secondary structure type per PDB.

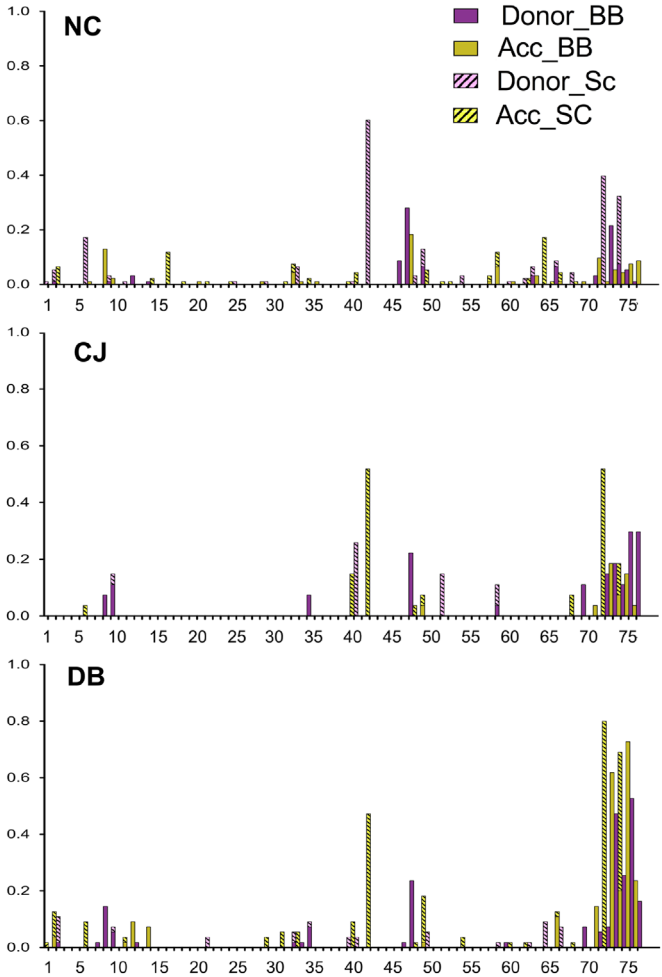

Fig. 5. Inter-molecular hydrogen bond sites on UBQ. We detected hydrogen bonds using Rosetta hydrogen bond score. Each hydrogen bond was classified as either a donor or acceptor and if the chemical moiety participating in the hydrogen bond belongs to the peptide backbone or the side chain. The ubiquitin structures are grouped by the following protein–protein interactions: NC: noncovalent, CJ: conjugated, DB: deubiquitinase. The *Y*-axis represents average number of hydrogen bonds per PDB. Redundant hydrogen bonds in structure containing multiple ubiquitin chains were only counted once per PDB.

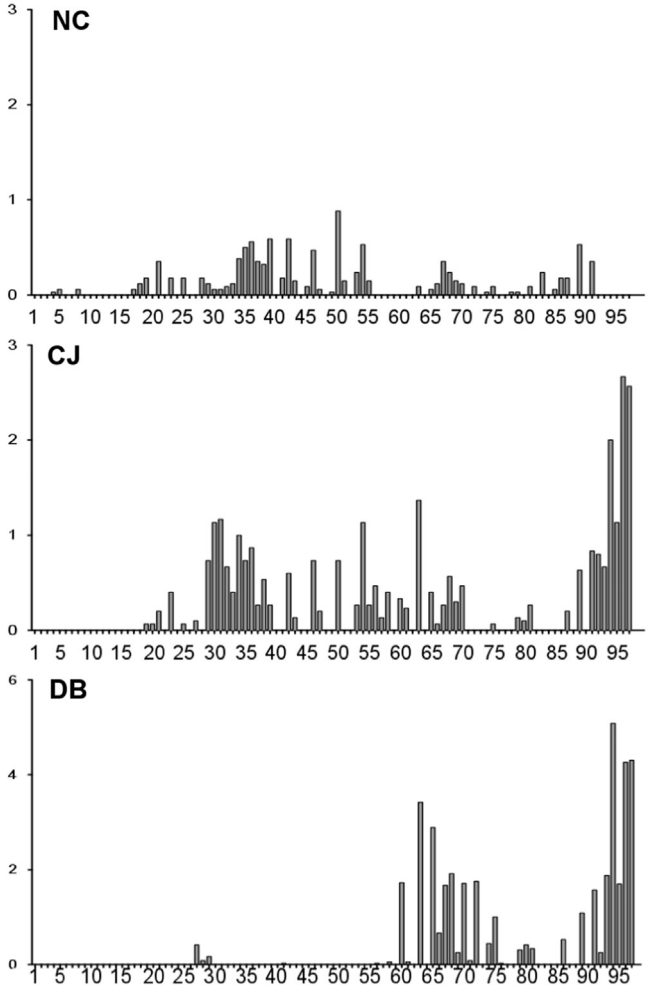

Fig. 6. Number of non-SUMO amino acid neighbors for each residue of SUMO using a 6 Å distance cutoff. The ubiquitin structures are grouped by the following protein–protein interactions: NC: noncovalent, CJ: conjugated, DB: deubiquitinase. The units on the *Y*-axis are average number of normalized non-UBQ neighboring residues per PDB. The SUMO1 numbering scheme is used for all SUMO molecules (Fig. 1).

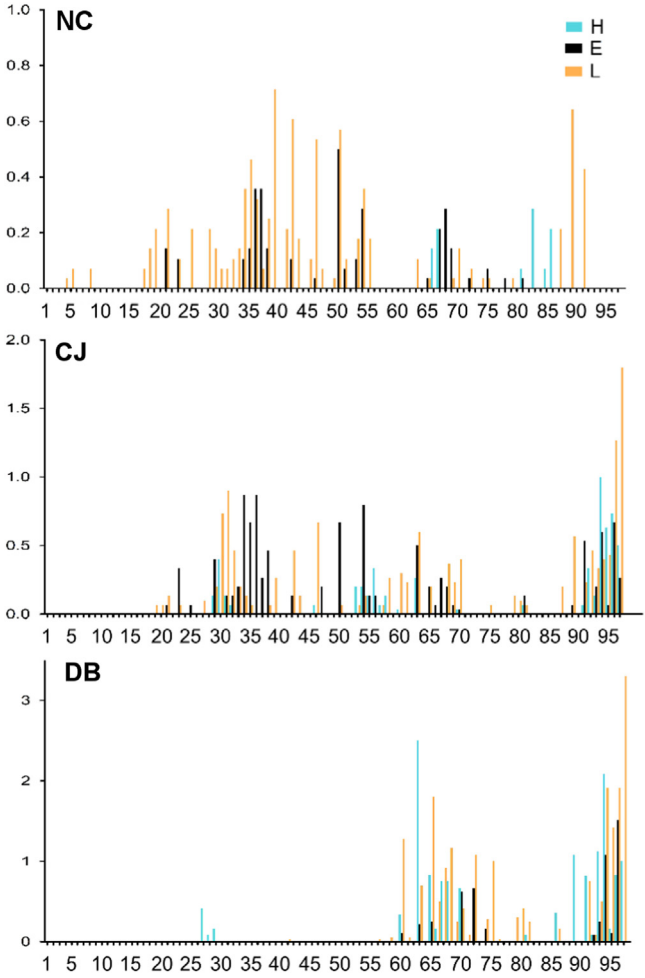

Fig. 7. Secondary structure of SUMO interacting motifs. The secondary structure was determined using the following simplified DSSP distinction, *H α-helix*, *E β-strand*, *L loop*. The SUMO structures are grouped by the following protein–protein interactions: NC: noncovalent, CJ: conjugated, DB: deubiquitinase. The *Y*-axis represents the average number of normalized interacting residues in each secondary structure element per PDB. The SUMO1 numbering scheme is used for all SUMO molecules (Fig. 1).

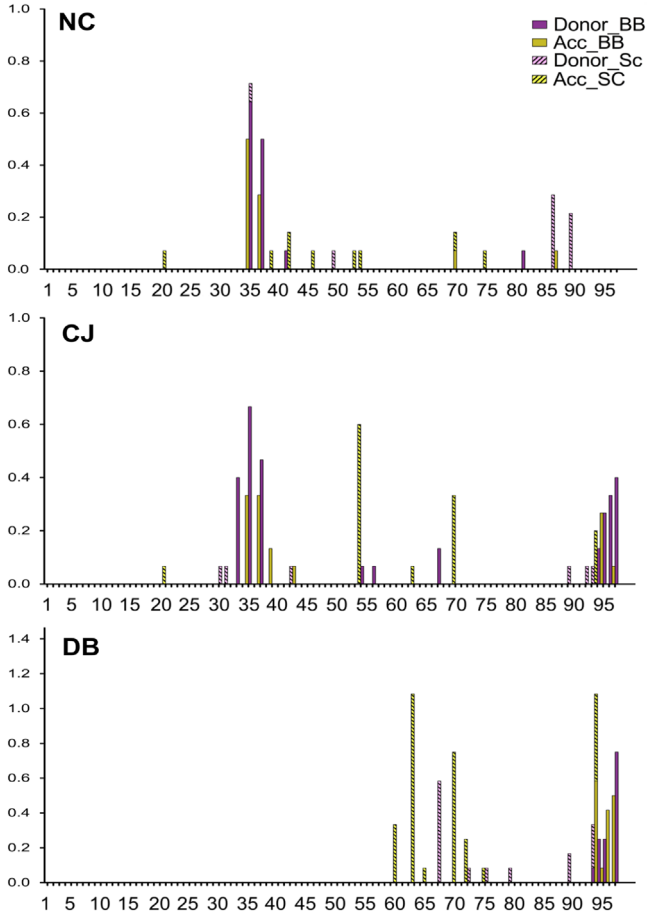

Fig. 8. Inter-molecular hydrogen bond sites on SUMO. We detected hydrogen bonds using Rosetta hydrogen bond score. Each hydrogen bond was classified as either a donor or acceptor and if the chemical moiety participating in the hydrogen bond belongs to the peptide backbone or the side chain. The SUMO structures are grouped by the following protein–protein interactions: NC: noncovalent, CJ: conjugated, DB: deubiquitinase. The *Y*-axis represents average number of hydrogen bonds per PDB. The SUMO1 numbering scheme is used for all SUMO molecules (Fig. 1). Redundant hydrogen bonds in structure containing multiple SUMO chains were only counted once per PDB.

found in Supporting file 3. To compute PDB averages for UBL recognition, each structure was normalized by the number of UBL chains participating in the same type of protein–protein interaction. Figs. 1–8

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2015.10.007.

References

[1] J.S. Harrison, T.M. Jacobs, K. Houlihan, K. Van Doorslaer, B. Kuhlman, UbSRD: The Ubiquitin Structural Relational Database, Journal of molecular biology (2015), http://dx.doi.org/10.1016/j.jmb.2015.09.011, Available at http://www.ncbi.nlm.nih.gov/pubmed/26392143.

[2] Z.-M. Zhang, S.B. Rothbart, D.F. Allison, Q. Cai, J.S. Harrison, L. Li, Y. Wang, B.D. Strahl, G.G. Wang, J. Song, An allosteric interaction links USP7 to deubiquitination and chromatin targeting of UHRF1, Cell Rep. 12 (9) (2015) 1400–1406. http://dx.doi.org/10.1016/j.celrep.2015.07.046, Available at⟨http://www.ncbi.nlm.nih.gov/pubmed/26299963⟩.

[3] S.B. Rothbart, B.M. Dickson, M.S. Ong, K. Krajewski, S. Houliston, D.B. Kireev, C.H. Arrowsmith, B.D. Strahl, Multivalent histone engagement by the linked tandem tudor and PHD domains of UHRF1 is required for the epigenetic inheritance of DNA methylation, Genes Dev. 27 (11) (2013) 1288–1298. http://dx.doi.org/10.1101/gad.220467.113, Available at⟨http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3690401&tool=pmcentrez&rendertype=abstract⟩.

[4] A.M. Burroughs, L.M. Iyer, L. Aravind, Structure and evolution of ubiquitin and ubiquitin-related domains, Methods Mol. Biol. (Clifton, N.J.) 832 (2012) 15–63. http://dx.doi.org/10.1007/978-1-61779-474-2_2, Available at⟨http://www.ncbi.nlm.nih.gov/pubmed/22350875⟩.

[5] T. Wauer, M. Simicek, A. Schubert, D. Komander, Mechanism of phospho-ubiquitin-induced PARKIN activation, Nature 524 (7565) (2015) 370–374. http://dx.doi.org/10.1038/nature14879, Available at⟨http://www.ncbi.nlm.nih.gov/pubmed/26161729⟩.

[6] A. Leaver-Fay, M.J. O'Meara, M. Tyka, R. Jacak, Y. Song, E.H. Kellogg, J. Thompson, I.W. Davis, R. a Pache, S. Lyskov, J.J. Gray, T. Kortemme, J.S. Richardson, J.J. Havranek, J. Snoeyink, D. Baker, B. Kuhlman, Scientific benchmarks for guiding macromolecular energy function improvement, Methods Enzymol. 523 (2013) 109–143. http://dx.doi.org/10.1016/B978-0-12-394292-0.00006-0, Available at⟨http://www.ncbi.nlm.nih.gov/pubmed/23422428⟩.

[7] M.J. O'Meara, A. Leaver-Fay, M.D. Tyka, A. Stein, K. Houlihan, F. DiMaio, P. Bradley, T. Kortemme, D. Baker, J. Snoeyink, B. Kuhlman, Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta, Journal of Chemical Theory and Computation 11 (2) (2015) 609–622. http://dx.doi.org/10.1021/ct500864r, Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4390092&tool=pmcentrez&rendertype=abstract.

[8] A. Leaver-fay, M. Tyka, S.M. Lewis, F. Lange, J. Thompson, R. Jacak, K. Kaufman, P.D. Renfrew, C.A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D.J. Mandell, F. Richter, Y.A. Ban, S.J. Fleishman, E. Corn, D.E. Kim, S. Lyskov, M. Berrondo, J. J. Havranek, S. Mentzer, Z. Popovic, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J.J. Gray, B. Kuhlman, D. Baker, P. Bradley, ROSETTA 3 : an object-oriented software suite for the simulation and design of macromolecules, Methods Enzymol. 487 (11) (2011) 545–574. http://dx.doi.org/10.1016/S0076-6879(11)87019-9.

[9] J. Pei, N.V. Grishin, PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information, Methods Mol. Biol. (Clifton, N.J.) 1079 (2014) 263–271. http://dx.doi.org/10.1007/978-1-62703-646-7_17, Available at⟨http://www.ncbi.nlm.nih.gov/pubmed/24170408⟩.

[10] G.M. Boratyn, A.A. Schäffer, R. Agarwala, S.F. Altschul, D.J. Lipman, T.L. Madden, Domain enhanced lookup time accelerated BLAST, Biol. Direct 7 (1) (2012) 12. http://dx.doi.org/10.1186/1745-6150-7-12, Available at⟨http://www.biology-direct.com/content/7/1/12⟩.

[11] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2, Available at⟨http://www.ncbi.nlm.nih.gov/pubmed/2231712⟩.

[12] S.J. Fleishman, A. Leaver-Fay, J.E. Corn, E.-M. Strauch, S.D. Khare, N. Koga, J. Ashworth, P. Murphy, F. Richter, G. Lemmon, J. Meiler, D. Baker, RosettaScripts: a scripting language interface to the rosetta macromolecular modeling suite, PloS One 6 (6) (2011) e20161. http://dx.doi.org/10.1371/journal.pone.0020161, Available at⟨http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3123292&tool=pmcentrez&rendertype=abstract⟩.

[13] "SQLite3 Syntax" Available at ⟨https://www.sqlite.org/lang.html⟩.